Data Mining

par Dr Ali LALOUCI
ali.lalouci@gmail.com
Département : Informatique
Institut : Mathématique et Informatique
1 Master I2A + STIC

https://www.univdocs.com/2020/05/data-mining-fouille-de-donnees.html

Présentation du Module

Intitulé du Master :

- **Intelligence Artificielle et ses Applications**
- **Sciences et Technologies de l'Information et de la Communication**

Intitulé de la matière : Data Mining

Semestre : 1

Intitulé de l'UE : UE Méthodologie (UEM1)

Connaissances préalables : Statistiques, Probabilités, Analyse de Données, Bases de Données

Présentation du Module

Crédits : 5

Coefficient : 3

1 Cours : Jeudi de 11:00 a 12:30

1 TD : Samedi de 15:30 a 17:00

1 TP: Lundi de 11:00 a 12:30

Note Module: Examen*0.6 + TD*0.2 + TP*0.2

Chapitre1

Introduction au Data Mining

Plan

1. Introduction
2. Définition de la fouille de données
3. Processus d'Extraction des connaissances
4. Typologies des méthodes de fouille de données
5. Prétraitement de données
6. Outils

1. Introduction

- La fouille de données ou data mining est une évolution naturelle dans l'exploitation des données par les être humains en utilisant les ordinateurs.
- On peut résumer cette évolution dans les points suivants :
  - ✓ **Début de l'informatique (années 40):** utilisation des ordinateurs pour les besoins de calcul.
    -Traitement *statistique* des données et *analyse de données* : prémices du Data Mining.
  - ✓ **Debut des années 60:** apparition du terme *Data Base* (base de données)

1. Introduction

- ✓ **Fin des années 80 :** exploitation du contenu des bases de données pour la recherche de règles d'association : utilisation du terme *database mining*.
- ✓ **1989 :** premier atelier sur la découverte de connaissances – proposition du terme *Knowledge Discovery* par Gregory Piatetsky-Shapiro .
- ✓ **1990 :** L'aide à la décision a introduit le traitement analytique en ligne (*OLAP*) : Online Analytical Processing
- ✓ **1995 :** première conférence sur le *Data Mining*.
- ✓ **2008:** la vague *NoSQL* introduit le Big Data, Graphe database,…

1. Introduction

À la croisée des chemins

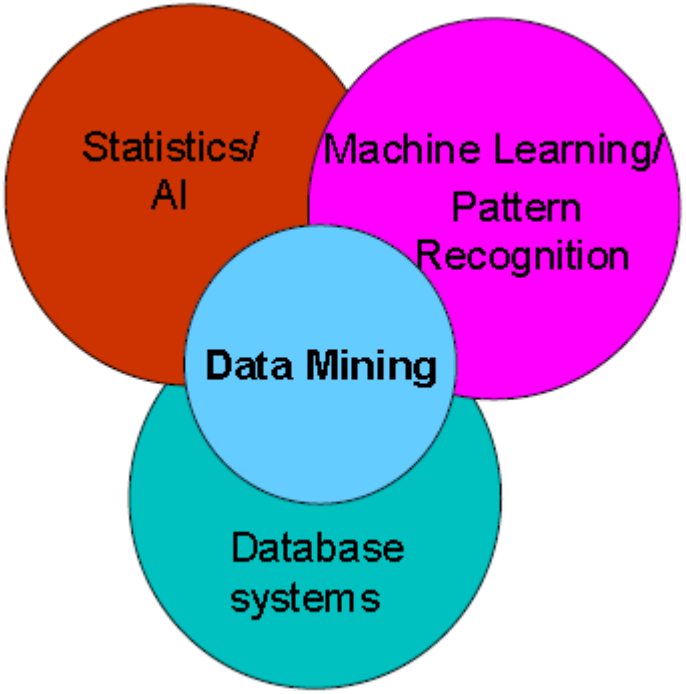


Figure 1. Origine du Data Mining.
(extraite de ''Tan et al, Introduction to Data Mining, 2004'')

2. Définition de la fouille de données

- Les définitions du data mining ne font pas parfois la différence entre le *Data Mining* et le *KDD* ou *Knowledge Discovery from Data* qu'on peut traduire par l'extraction des connaissances à partir des données.

Définition 1: « l'extraction de connaissances à partir des données est un processus non trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données » **Fayyad**

Définition 2: « l'extraction non triviale d'informations implicite, précédemment inconnue, et potentiellement utiles à partir des données ». **Frawley**

2. Définition de la fouille de données

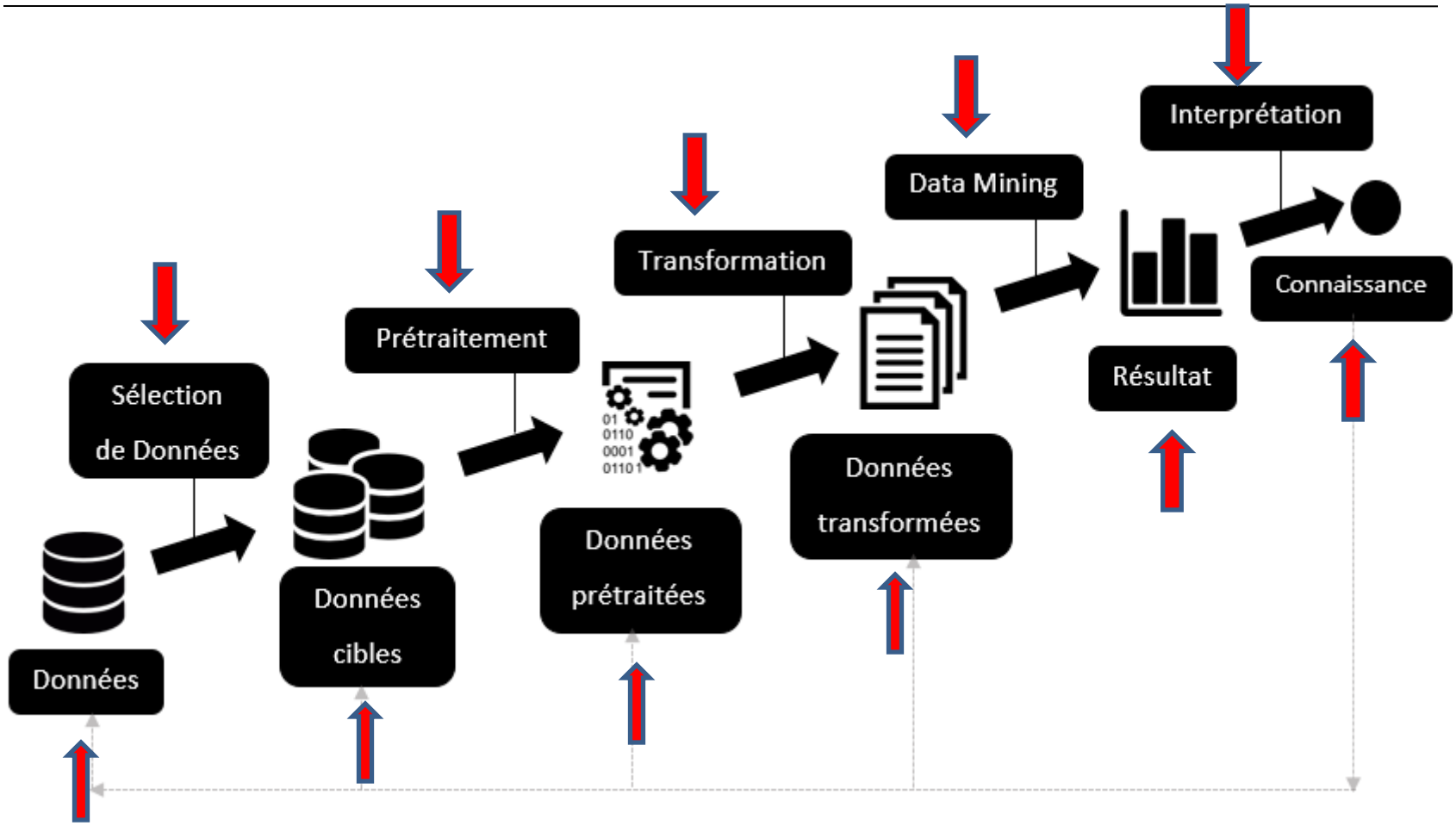
Définition 3 : « Un Processus inductif, *itératif et interactif* de découverte dans les BD larges de modèles de données *valides, nouveaux, utiles et compréhensibles* ». *générale*

- ✓ *Itératif* : nécessite plusieurs passes
- ✓ *Interactif* : l'utilisateur est dans la boucle du processus
- ✓ *Valides* : valables dans le futur
- ✓ *Nouveaux* : non prévisibles
- ✓ *Utiles* : permettent à l'utilisateur de prendre des décisions
- ✓ *Compréhensibles* : présentation simple.

Domaines d'Application

- Log
  spe
- Do
  co

3. Processus d'Extraction des connaissances



3. Processus d'Extraction des connaissances

- Les étapes précédentes d'un processus ECD peuvent se résumer en trois grandes phases:

3. Processus d'Extraction des connaissances

1. Prétraitement (ou préparation des données)

C'est une étape essentielle, car les données brutes sont souvent *incomplètes, bruitées ou incohérentes*. Elle comprend généralement :

- Collecte des données (depuis diverses sources)
- Nettoyage des données (suppression des doublons, valeurs manquantes, correction d'erreurs)
- Transformation / normalisation (mise à l'échelle, codage, agrégation, etc.)
- Sélection des variables pertinentes (feature selection), faire des échantillons représentatifs de la population d'origine.

👉 Objectif : obtenir un *jeu de données propre et exploitable* pour la phase suivante.

3. Processus d'Extraction des connaissances

2. Phase de Data Mining

C'est le cœur du processus, où les techniques d'analyse et d'extraction de connaissances sont appliquées. Selon les objectifs, on peut utiliser :

- Classification (ex. : arbres de décision, SVM, réseaux de neurones)
- Clustering (ex. : k-means, DBSCAN)
- Association (ex. : règles d'association de type Apriori)
- Régression, etc.

👉 Objectif : extraire des modèles ou des motifs significatifs à partir des données préparées.

3. Processus d'Extraction des connaissances

3. Post-traitement (interprétation et évaluation)

Après avoir obtenu les modèles, il faut :

- Évaluer la qualité et la pertinence des résultats (mesures de précision, de rappel, etc.)
- Interpréter les modèles extraits pour en tirer des connaissances utiles
- Visualiser les résultats (graphiques, rapports, tableaux de bord), indicateurs statistiques, ,,,
- Mettre en œuvre les connaissances dans un système de décision

👉 Objectif : transformer les résultats du Data Mining en connaissances exploitables pour la prise de décision.

4. Typologie de Méthodes

Selon les d'objectifs :

- Méthodes de Classification :

L'objectif de la classification consiste à examiner les caractéristiques d'un objet (ou individu) afin de lui attribuer une *classe* ou une *catégorie prédéfinie*.

✓ Exemples :

- Décider d'accorder ou non un prêt à un client.
- Établir un diagnostic médical à partir de symptômes observés.

✓ Algorithmes couramment utilisés :

- Arbres de décision
- Machines à vecteurs de support (SVM)
- Classifieur bayésien (Naïve Bayes)

4. Typologie de Méthodes

Selon les d'objectifs :

- Méthodes de Prédiction :

L'objectif de la prédiction consiste à estimer la valeur future ou inconnue d'un attribut à partir d'autres attributs connus.

✓ Exemples :

- Prédire la qualité d'un client en fonction de son revenu, de son âge ou de son nombre d'enfants.
- Prédire le prix d'un appartement en fonction de sa surface, de son étage, ou de sa localisation.

✓ Algorithmes couramment utilisés :

- Régression linéaire
- (On peut aussi citer : régression logistique, réseaux de neurones, arbres de décision, etc., selon le type de donnée et de sortie attendue.)

4. Typologie de Méthodes

Selon Types d'objectifs :

- Association :

L'objectif de l'analyse d'association consiste à identifier les relations ou corrélations entre différents attributs dans un ensemble de données.

✓ Exemples :

- Découvrir que « les clients qui achètent du pain achètent souvent aussi du beurre ».
- Trouver des relations entre produits dans les ventes d'un supermarché.

✓ Algorithmes couramment utilisés :

- Règles d'association (ex. : algorithme *Apriori*, *FP-Growth*)
- Analyse de corrélation entre attributs

4. Typologie de Méthodes

Selon Types d'objectifs :

- Segmentation (ou Clustering) :

L'objectif de la segmentation est de former des groupes homogènes d'individus ou d'objets ayant des caractéristiques similaires, à l'intérieur d'une population donnée.

✓ Exemples :

- Segmenter les clients d'une entreprise selon leur comportement d'achat.
- Regrouper des articles ou documents similaires selon leur contenu.

✓ Algorithmes couramment utilisés :

- K-Means
- Analyse Hiérarchique de Clustering (AHC)

4. Typologie de Methodes

Types de modèles obtenus :

- Modèles prédictifs :
  - ✓ utilisent les données avec des résultats connus pour développer des modèles permettant de prédire les valeurs d'autres données
  - ✓ Ex: modèle permettant de prédire les clients qui ne rembourseront pas leur crédit
  - ✓ Algorithemes : arbres de decision, Regression ,…
- Modèles descriptifs :
  - ✓ proposent des descriptions des données pour aider à la prise de décision
  - ✓ aident à la construction de modèles prédictifs
  - ✓ Algorithemes : - Clustring, Régles d'association,…

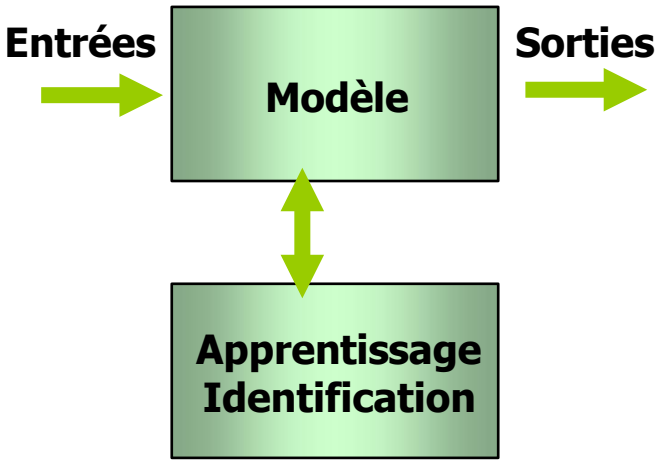
4. Typologie de Methodes

Types d'apprentissage

- Apprentissage supervisé - Fouille supervisée :
  - ✓ processus dans lequel l'apprenant reçoit des exemples d'apprentissage comprenant à la fois des données d'entrée et de sortie
  - ✓ les exemples d'apprentissage sont fournis avec leur classe (valeur de sortie prédite)

  - ✓ But : classer correctement un nouvel exemple (généralisation)
  - ✓ utilisées principalement:

    en classification et prédiction
  - ✓ Algorithmes : classifications, régression



4. Typologie de Methodes

Types d'apprentissage

- Apprentissage non supervisé - Fouille non supervisée :

✓ processus dans lequel l'apprenant reçoit desexemples d'apprentissage ne comprenant que des données d'entrée,

✓ pas de notion de classe

✓ But : regrouper les exemples en « paquets » (clusters) d'exemples similaires (on peut ensuite donner un nom à chaque cluster)

✓ utilisé principalement en association et segmentation

✓ Algorithmes : Segmentation,regroupement, découverte d'associations et de règles

5. Prétraitement de données

Définition de la donnée

- *Enregistrement* au sens des bases de données (*individu* en statistiques, *instance* en terminologie orienté objet ou *tuple* du point de vue base de données)
- Un point dans un espace euclidien, ou un vecteur dans un espace vectoriel
- Une donnée est caractérisée par un ensemble *d'attributs* (variable, caractéristique,…) qui peut être *qualitatif* ou *quantitatif*
- Souvent, les données exploitées par le DM doivent avoir une forme tabulaire (ligne/colonne). Si les données n'ont pas cette forme, elles sont transformées et adaptées

5. Prétraitement de données

Why Data Pre-processing?

- Dans le monde réel, les données sont sales.
  - ✓ Incomplètes : valeurs d'attribut manquantes, certains attributs d'intérêt manquants, ou ne contenant que des données agrégées.
  - ✓ Bruyantes : erreurs ou valeurs aberrantes.
  - ✓ Incohérentes : divergences de codes ou de noms.
- Pas de données de qualité, pas de résultats d'exploration de données de qualité !
  - ✓ Les décisions de qualité doivent être fondées sur des données de qualité.
  - ✓ Les données nécessitent une intégration cohérente,

5. Prétraitement de données

Nettoyage des données

- Consiste à traiter les *valeurs manquantes*, les *données bruyantes*, …
- identifier ou supprimer les *valeurs aberrantes* et Corriger les *données incohérentes*.
- On utilise différentes techniques comme :
  - ✓ *le remplacement* de la valeur absente par la valeur la plus fréquente,
  - ✓ *l'estimation* de la valeur absente à partir des valeurs existantes, etc

5. Prétraitement de données

Les valeurs manquantes

- Les données manquantes peuvent être dues à :
  - ✓ dysfonctionnement de l'équipement
  - ✓ incohérentes avec d'autres données enregistrées et donc supprimées
  - ✓ données non saisies en raison d'un malentendu
  - ✓ certaines données peuvent ne pas être considérées comme importantes au moment de la saisie
  - ✓ pas d'historique de registre ni de modification des données

5. Prétraitement de données

Les valeurs manquantes

- Comment gérer les données manquantes ?

✓ Ignorer le tuple : pas efficace lorsque le pourcentage de valeurs manquantes par attribut varie considérablement

✓ Remplir manuellement la valeur manquante : fastidieux + irréalisable ?

✓ Utiliser une constante globale pour remplir la valeur manquante : par exemple, "inconnu", une nouvelle classe ? !

✓ Utiliser l'attribut moyen pour remplir la valeur manquante

✓ Utiliser l'attribut moyen pour tous les échantillons appartenant à la même classe pour remplir la valeur manquante : smarter

✓ Utiliser la valeur la plus probable pour remplir la valeur manquante : *basée sur l'inférence, telle que la formule bayésienne ou l'arbre de décision*

6. Outils du Data mining

- R, SAS, Python ou encore Matlab sont les principaux langages interprétés d'analyse de données et du data mining. offrent également une très large bibliothèque de fonctions statistiques extensibles avec des packages grâce auxquels on peut faire appel à des algorithmes déjà implémentés.

- Logiciels d'analyse statistique avec une interface simple tels que Oracle E-Business Suite, Watson (édité par IBM), SAS BI, KNIME, RevolutionR de Revolution Analytics, Tableau ou encore DataIKU

.

6. Outils du Data mining

- Logiciels commercialisés :
  - S-PLUSTM de Insight,
  - AliceTM de Isoft,
  - Predict TM de Neuralware,
- Logiciels gratuits :
  - Weka
  - Tanagra
  - Orange
  - R (version gratuite de S-PLUS)